

C02029: Doctor of Philosophy
CRICOS Code: 009469A
32903 PhD Thesis: Analytics
January 2019

A Study on
Automated Handwriting Understanding

Chandranath Adak

School of Software
Faculty of Engg. & IT
University of Technology Sydney
NSW - 2007, Australia

A Study on Automated Handwriting Understanding

*A thesis submitted in partial fulfilment of the requirements
for the degree of*

Doctor of Philosophy

in

Analytics

by

Chandranath Adak

to

School of Software

Faculty of Engineering and Information Technology

University of Technology Sydney

NSW - 2007, Australia

January 2019

ABSTRACT

*H*andwriting is a concatenation of graphical symbols drawn by a pen or other writing instruments, using a hand in order to represent linguistic constructs for communication and knowledge storage. These graphical marks/writing symbols have deep orthographic relation to the phonology of a spoken language. However, to a machine, handwriting is nothing but a pattern. Therefore, recognition of this pattern is performed in order to read a manuscript by a computer. Such a process of automatic character pattern recognition from an optically scanned document image is called OCR (Optical Character Recognition). Nowadays, the computer vision method is not limited to simply recognizing patterns/objects. It tries to endow the machine a human-like intelligent ability. The main goal of this research is using computer vision to bridge the gap between pattern recognition and human perception of handwriting. In this thesis, we focus on understanding the handwriting, which is beyond simply recognizing the characters by OCR. Towards this aim, we peek into the implicit information of handwriting to understand some inherent characteristics.

In this thesis, we concentrate on three aspects. First, understanding the generation of handwritten information by the writing body; second, understanding the writing strokes in regards to the quality of handwriting; third, understanding the content revealing handwritten word entities. Thus, the thesis contains three parts. Regardless of past researches on writer inspection, it is hard to find an empirical study performed on intra-variable handwriting, although such variation should be an important concern. The first part of this thesis addresses this concern. Besides, this part inspects the writer on some unconventional aspects, e.g., writing variability over struck-out texts, multiple scripts, etc. The second part makes a pioneering contribution to understanding writing stroke information in multiple facets, such as legibility, aesthetics, difficulty, and idiosyncrasy of strokes. The third part of the thesis approaches to comprehend the content of the handwritten document using computer vision, without the aid of a transcription engine or the natural language processing which, according to our knowledge, is the earliest attempt of its kind.

This research has adapted the traditional machine learning approaches as well as state-of-the-art deep learning approaches and has proposed new techniques to automate the process of handwriting understanding. The performed experiments have produced encouraging results, which ensure the applicability of the proposed research. This study has an impact on general image processing, pattern recognition, machine learning, and deep

learning domains, especially on document image processing and handwriting processing. Moreover, this research contributes to forensics for questioned document examination, biometrics for behavioral analysis through handwriting, library science/archival science for e-archiving of the manuscript, and data science. According to us, this study has pushed the frontiers of handwriting-related research.

AUTHOR'S DECLARATION

I, *Chandranath Adak* declare that this thesis, submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Software, Faculty of Engineering and Information Technology at the University of Technology Sydney, Australia, is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Production Note:

SIGNATURE: Signature removed prior to publication.

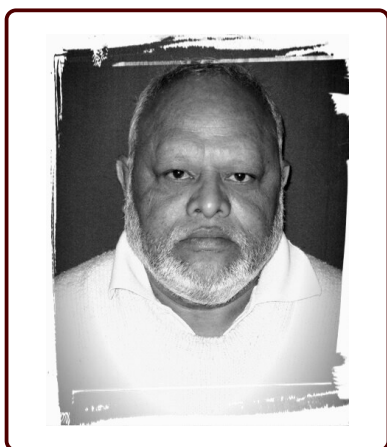
[Chandranath Adak]

DATE: 18th January, 2019

PLACE: Sydney, Australia

DEDICATION

To my father ...



Late Baidyanath Adak

ACKNOWLEDGMENTS

My sincere gratitude is to my principal supervisor Prof. Michael Blumenstein, University of Technology Sydney (UTS), for his guidance and support during my Ph.D. candidature. His continuous encouragement and advice have made our research very productive. I would also like to convey my gratitude to my external co-supervisor Prof. Bidyut Baran Chaudhuri, Indian Statistical Institute (ISI), for his guidance, advice and constant motivation. For the advancement of our research, he literally acted as a critical reviewer with counter arguments and meticulous insistence on various aspects.

I would like to thank my co-supervisor Prof. Chin-Teng Lin, UTS, for his support and advice. I extend my special thanks to A/Prof. Simone Marinai, DINFO, University of Florence, Italy for his insightful advice during my visit to his lab, and IEEE Computational Intelligence Society for providing research grant for this visit.

I transferred my Ph.D. from Griffith University, Australia to UTS, following my principal supervisor Prof. Blumenstein. I heartily thank School of ICT, IIIS, and GGRS of Griffith University for supporting my initial Ph.D. research. I gratefully acknowledge all the faculty members, admin staffs, fellow researchers, GUGC student guild members of Griffith University. A special thanks to then my associate supervisor A/Prof. Jun Jo, Griffith University, for his constant support during my Griffith days.

Moreover, I express my gratitude to all the faculty members, admin staffs, fellow researchers of School of Software, CAI, and FEIT; and GRS staffs of UTS for their continuous support. I wish to thank all the faculty members and fellow researchers of CVPR Unit, Indian Statistical Institute for fruitful discussion during my ISI visits.

I also benefited from the constructive comments and suggestions made by several fellow researchers, domain experts met in conferences/academic gatherings, and also from anonymous reviewers of my papers. I thank all of them.

Heartfelt thanks to Dr. Abhijit Das (INRIA), Dr. Soumya Dutta (LANL), Tanmoy Nandi (ISI), Sumit Kr. Saha (ISI), Muhammad Saqib (UTS), Di Wu (UTS) for their help and support.

Finally, I am thankful to all my family members, especially my wife Soumi Chattopadhyay (IIITG), for constant encouragement, enthusiasm and support. I thank all those, whom I missed out from the above list.

LIST OF PUBLICATIONS

RELATED TO THE THESIS :

1. C. ADAK, B. B. CHAUDHURI, M. BLUMENSTEIN, *An Empirical Study on Writer Identification & Verification from Intra-variable Individual Handwriting*, **IEEE Access**, vol. 7, no. 1, pp. 24738-24758, 2019.
2. C. ADAK, B. B. CHAUDHURI, M. BLUMENSTEIN, *A Study on Idiosyncratic Handwriting with Impact on Writer Identification*, Proc. 16th Int. Conf. on Frontiers in Handwriting Recognition (**ICFHR**), pp. 193-198, Niagara Falls, USA, 5-8 Aug., 2018.
3. C. ADAK, B. B. CHAUDHURI, M. BLUMENSTEIN, *Cognitive Analysis for Reading and Writing of Bengali Conjuncts*, Proc. 31st Int. Joint Conference on Neural Networks (**IJCNN**), Rio de Janeiro, Brazil, 8-13 July, 2018.
4. C. ADAK, S. MARINAI, B. B. CHAUDHURI, M. BLUMENSTEIN, *Offline Bengali Writer Verification by PDF-CNN and Siamese Net*, Proc. 13th Int. Workshop on Document Analysis Systems (**DAS**), pp. 381-386, Austria, 24-27 Apr., 2018.
5. C. ADAK, B. B. CHAUDHURI, M. BLUMENSTEIN, *Legibility and Aesthetic Analysis of Handwriting*, Proc. 14th Int. Conf. on Document Analysis and Recognition (**ICDAR**), pp. 175-182, Kyoto, Japan, 9-15 Nov., 2017.
6. C. ADAK, B. B. CHAUDHURI, M. BLUMENSTEIN, *Impact of Struck-out Text on Writer Identification*, Proc. 30th Int. Joint Conf. on Neural Networks (**IJCNN**), pp. 1465-1471, Anchorage, Alaska, USA, 14-19 May, 2017.
7. C. ADAK, B. B. CHAUDHURI, M. BLUMENSTEIN, *Writer Identification by Training on One Script but Testing on Another*, Proc. 23rd Int. Conference on Pattern Recognition (**ICPR**), pp. 1148-1153, Cancun, Mexico, 4-8 Dec., 2016.

-
8. C. ADAK, B. B. CHAUDHURI, M. BLUMENSTEIN, *Named Entity Recognition from Unstructured Handwritten Document Images*, Proc. 12th Int. Workshop on Document Analysis Systems (**DAS**), pp.375-380, Santorini, Greece, 11-14 Apr., 2016.

OTHERS :

9. C. ADAK, B. B. CHAUDHURI, M. BLUMENSTEIN, *Offline Cursive Bengali Word Recognition using CNNs with a Recurrent Model*, Proc. 15th Int. Conf. on Frontiers in Handwriting Recognition (**ICFHR**), pp. 429-434, Shenzhen, China, 23-26 Oct., 2016.
10. C. ADAK, P. MAITRA, B. B. CHAUDHURI, M. BLUMENSTEIN, *Binarization of Old Halftone Text Documents*, Proc. **IEEE TENCON**, IEEE Conf. # 35439, pp. 1-5, Macau, China, 1-4 Nov., 2015.
11. B. B. CHAUDHURI[†], C. ADAK[†], *An Approach for Detecting and Cleaning of Struck-out Handwritten Text*, **Pattern Recognition**, vol. 61, pp. 282-294, January 2017.
- [†]*Joint first author.*

TABLE OF CONTENTS

	Page
Abstract	i
List of Publications	ix
List of Figures	xvii
List of Tables	xxi
1 Introduction	1
1.1 Handwriting Understanding	5
1.2 Motivation	7
1.3 Contributions	8
1.4 Related Literature	10
1.4.1 Writer Investigation	10
1.4.2 Writing Stroke Understanding	15
1.4.3 Content Understanding	15
1.4.4 Handwriting Database	16
1.5 Organization of the Thesis	17
I Understanding the Writer	19
2 Writer Verification by <i>PDF-CNN</i> & Siamese Net	21
2.1 Proposed Method	23
2.1.1 Handcrafted Feature PDF	23
2.1.2 Handcrafted Feature PDF Hybridized with Auto-derived CNN Features (PDF-CNN)	25
2.1.3 Classification / Writer Verification	26

TABLE OF CONTENTS

2.2	Experiments and Discussion	28
2.2.1	Database Employed	28
2.2.2	Results and Evaluation	28
2.2.3	Comparison	31
2.3	Summary	32
3	Writer Identification across Scripts	33
3.1	Proposed Method	36
3.1.1	Feature Extraction	36
3.1.2	Classification	40
3.2	Experiments and Discussion	41
3.2.1	Database Employed	41
3.2.2	Results and Evaluation	42
3.3	Summary	46
4	Writer Identification and Verification from Intra-variable Writing	47
4.1	Related Work on Intra-variable Handwriting	51
4.2	Experimental Dataset	52
4.2.1	Controlled Database (D_c)	52
4.2.2	Uncontrolled Database (D_{uc})	54
4.3	Preprocessing	57
4.4	Handcrafted Feature Extraction	57
4.4.1	Macro-Micro Features (F_{MM})	57
4.4.2	Contour Direction and Hinge Features (F_{DH}):	58
4.4.3	Direction and Curvature Features at Keypoints (F_{DC})	59
4.5	Auto-derived Feature Extraction	60
4.5.1	Basic_CNN	60
4.5.2	SqueezeNet	61
4.5.3	GoogLeNet	62
4.5.4	Xception Net	62
4.5.5	VGG-16	63
4.5.6	ResNet-101	63
4.6	Writer Identification	64
4.6.1	Handcrafted Feature-based Identification	64
4.6.2	Auto-derived Feature-based Identification	64
4.7	Writer Verification	66

4.7.1	Handcrafted Feature-based Verification	66
4.7.2	Auto-derived Feature-based Verification	67
4.8	Experiments and Discussion	69
4.8.1	Database Employed	69
4.8.2	Writer Identification Performance	71
4.8.3	Writer Verification Performance	77
4.8.4	Observations	80
4.8.5	Writer Identification/Verification by Pre-training	82
4.8.6	Writer Identification/Verification on Enlarged Writer Set	84
4.8.7	Comparison	85
4.9	Summary	86
5	Impact of Struck-out Text on Writer Identification	87
5.1	Proposed Method	89
5.1.1	Normal and Struck-out Text Separation	90
5.1.2	Writer Identification	91
5.2	Experiments and Discussion	93
5.2.1	Database Employed	93
5.2.2	Results and Evaluation	94
5.2.3	Comparison	97
5.3	Summary	99
II	Understanding Writing Strokes	101
6	Legibility and Aesthetic Analysis of Writing Strokes	103
6.1	Proposed Method	108
6.1.1	Problem Formulation	108
6.1.2	Preprocessing	108
6.1.3	Legibility Analysis	109
6.1.4	Aesthetic Analysis	110
6.2	Experiments and Discussion	113
6.2.1	Database Employed	113
6.2.2	Results and Evaluation	115
6.2.3	Comparison	118
6.3	Summary	119

TABLE OF CONTENTS

7	Difficulty Analysis for Writing and Reading of Character Strokes	121
7.1	Briefings on Bengali Script	122
7.2	Proposed Method	123
7.2.1	Problem Formulation	124
7.2.2	Auto-derived Feature-based Inception Network	124
7.2.3	Hand-crafted Feature-based SVM	126
7.3	Experiments and Discussion	129
7.3.1	Database Employed	129
7.3.2	Results and Evaluation	131
7.3.3	Correlation Testing of Conjunct Character Reading / Writing Diffi- culty Analysis	133
7.3.4	Comparison	134
7.4	Summary	134
8	Idiosyncratic Writing Stroke Analysis and Impact on Writer Identifi- cation	135
8.1	Proposed Method	138
8.1.1	Idiosyncrasy Analysis	138
8.1.2	Writer Identification	141
8.2	Experiments and Discussion	142
8.2.1	Database Employed	142
8.2.2	Results and Evaluation	143
8.2.3	Comparison	146
8.3	Summary	146
III	Understanding the Content	149
9	Content Revealing Word Recognition from Manuscript Images	151
9.1	Proposed Method	153
9.1.1	Binarization	154
9.1.2	Word Segmentation	154
9.1.3	Slant / Skew / Baseline Correction	155
9.1.4	Characteristics of Named Entities	155
9.1.5	Feature Extraction	157
9.1.6	Classification	159

9.1.7	Post-processing	160
9.2	Experiments and Discussion	161
9.2.1	Database Employed	161
9.2.2	Results and Evaluation	161
9.3	Summary	164
10	Conclusions and Future Scope	165
	Bibliography	169

LIST OF FIGURES

FIGURE	Page
1.1 Some examples of handwriting availability in the document sample: (a) letter, (b) printed form entry, (c) signature, (d) boxed-space form entry, (e) hand-note on a typed letter, (f) writing merged with ruled-line and seal, (g) writing on a ruled-page and mixed with doodle/drawing.	4
1.2 Handwriting variations: (a),(b),(c): low inter-variability although written by 3 different writers; (d),(e): high intra-variability although written by the same writer.	5
1.3 Illumination of information in handwriting.	6
2.1 Every pair in a box comprises the same compound Bengali character written in two different styles [44].	22
2.2 PDF-CNN model diagram.	24
2.3 (a) Textural PDF-CNN and (b) Allographic PDF-CNN architectures.	26
2.4 Textural CNN-MLP performance evaluation: FAR and FRR plot. EER = 3.42.	29
2.5 Allographic CNN-MLP performance evaluation: FAR and FRR plot. EER =	29
2.6 Textural CNN-Siamese performance evaluation: FAR and FRR plot. EER = 2.36.	30
2.7 Allographic CNN-Siamese performance evaluation: FAR and FRR plot. EER = 3.51.	30
3.1 Handwriting specimens of two separate writers in both (a),(c) English and (b),(d) Bengali script.	34
3.2 Image samples from our database of (a)-(b) Writer-001 and (c)-(d) Writer-064. (a), (c) Bengali and (b), (d) English handwritings.	42
4.1 Ideal writer identification and verification system.	48

4.2	Intra-variable Bengali handwritten samples, <i>left column</i> : (A1), (A2), (A3) samples are of Writer-A, <i>right column</i> : (B1), (B2), (B3) samples are of Writer-B. The intra-variability of Writer-A's samples is less compared to Writer-B's samples, as confirmed by handwriting experts.	49
4.3	BCNN _{char} : Basic_CNN architecture as a feature extractor while using patch _{char}	61
4.4	(a) patch _{char} of size $n_{char} \times n_{char}$, (b) patch _{allo} of size $n_{allo} \times n_{allo}$ is shown in dark-gray, and the zero-padding, bounding the patch _{allo} is shown in light-gray color.	62
4.5	Writer identification strategies: (a) <i>Strategy-Major</i> , (b) <i>Strategy-Mean</i>	66
4.6	Siamese architecture.	68
4.7	Pictorial representation of a Bengali character component: keypoints are marked by red dots and the six edges are numbered from '1' to '6'.	70
4.8	Augmented text samples, generated from 2 handwritten pages of a writer in set S_f . The subset S_{f1} contains green colored 22 samples for training, S_{f2} contains blue colored 11 samples for validation, and S_{f3} contains orange colored 11 samples for testing.	71
4.9	Radar plot of Top-1 writer identification performance using handcrafted features. <i>Left</i> : representing Table 4.5 on D_c and <i>Right</i> : representing Table 4.6 on D_{uc}	81
4.10	Radar plot of Top-1 writer identification performance using auto-derived features. <i>Left</i> : representing Table 4.7 on D_c and <i>Right</i> : representing Table 4.8 on D_{uc}	82
4.11	Radar plot of writer verification performance using handcrafted features. <i>Left</i> : representing Table 4.9 on D_c and <i>Right</i> : representing Table 4.10 on D_{uc}	82
4.12	Radar plot of writer verification performance using auto-derived features. <i>Left</i> : representing Table 4.11 on D_c and <i>Right</i> : representing Table 4.12 on D_{uc}	83
5.1	Examples of (a) English and (b) Bengali handwritten documents containing struck-out texts, marked by red boxes.	88
5.2	Work-flow of the proposed method.	89
5.3	Our CNN architecture as a feature extractor.	90
6.1	Examples of (a) English and (b) Bengali readable texts having character-formation ambiguity.	104
6.2	Examples of (a) high and (b) low aesthetic handwritten Bengali document.	106
6.3	Our CNN architecture as a feature extractor.	109

6.4	Aesthetic analysis with subjective scores of G_{Ben} and G_{nonBen} groups.	118
6.5	The impact of different hand-crafted features for aesthetic analysis, while testing on D_T'' after training by entire training set.	119
7.1	Printed Bengali conjunct characters: 1 st row shows conjoining basic characters, 2 nd row depicts transparent conjuncts (<i>Bangla Akademi</i> font), 3 rd row represents non-transparent conjuncts (<i>AdorshoLipi</i> font). Each column is comprised of the same conjunct.	123
7.2	Handwritten Bengali conjunct characters: Each box contains a pair of the same conjunct, written in two different ways.	123
7.3	Inception Module.	125
7.4	Our adapted Inception Network model.	125
7.5	(a) Concave water reservoirs: Top, Bottom, Left, and Right reservoirs are shown by the color blue, magenta, red, and green, respectively. (b) Effort measure in 8-directions from a central pixel.	128
7.6	Barcode representation of reading/writing difficulty analysis: R_A and W_A denote Reading and Writing analysis using Auto-derived features, R_H and W_H represent Reading and Writing analysis using Hand-crafted features, respectively.	133
8.1	Examples of highly idiosyncratic handwritten character in (a) English and (b) Bengali script, enclosed in red colored boxes.	136
8.2	Two idiosyncratic characters marked by green and blue dotted boxes.	136
8.3	Highly idiosyncratic text-patches from <i>R. N. Tagore's</i> manuscript.	136
8.4	Inception Module (IM).	140
8.5	Our adapted Inception Network.	140
8.6	Fire module.	141
8.7	Our adapted SqueezeNet architecture.	142
9.1	Content-revealing named entities of person, organization and year, marked by the red, blue and green color boxes, respectively.	152
9.2	Proposed framework.	154
9.3	Qualitative result: NE identification in document images of (a) GWdb, (b) QSAdb, (c) IAMdb. Color boxes denote true positive (red), false negative (blue) and false positive (green) cases.	162

LIST OF TABLES

TABLE	Page
1.1 Some popular offline handwriting databases	16
2.1 Writer verification performance	31
2.2 Comparison with some state-of-the-art methods	32
3.1 Top-1 writer identification performance	44
3.2 Top-2 writer identification performance	44
3.3 Top-5 writer identification performance	45
3.4 Writer verification performance	45
4.1 Clustering method evaluation on D_c	56
4.2 Top-1 writer identification performance of system $WI_{D_c}F_{MM}$	72
4.3 Top-1 writer identification performance of system $WI_{D_c}F_{DH}$	72
4.4 Top-1 writer identification performance of system $WI_{D_c}F_{DC}$	73
4.5 Top-1 writer identification performance using handcrafted features on D_c . .	74
4.6 Top-1 writer identification performance using handcrafted features on D_{uc} .	74
4.7 Top-1 writer identification performance using auto-derived features on D_c . .	76
4.8 Top-1 writer identification performance using auto-derived features on D_{uc} .	77
4.9 Writer verification performance using handcrafted features on D_c	78
4.10 Writer verification performance using handcrafted features on D_{uc}	78
4.11 Writer verification performance using auto-derived features on D_c	79
4.12 Writer verification performance using auto-derived features on D_{uc}	80
4.13 Top-1 writer identification by pre-training	83
4.14 Writer verification by pre-training	84
4.15 Top-1 writer identification on enlarged writer set	85
4.16 Top-1 writer verification on an enlarged writer set	85
5.1 Struck-out text detection performance	95

5.2	Writer identification performance using SVM	96
5.3	Writer identification performance using CNN-RNN	97
5.4	Comparison of struck-out text detection	98
6.1	Legibility analysis	116
6.2	Aesthetic analysis by G_{Ben}	117
6.3	Aesthetic analysis by G_{nonBen}	117
7.1	Auto-derived feature-based analysis	132
7.2	Hand-crafted feature-based analysis	132
8.1	Idiosyncratic handwriting analysis	144
8.2	Writer identification performance	145
9.1	Evaluation of named entity identification	163
9.2	Impact of employed features	163